Zoosymposia 5: 331–337 (2011)
www.mapress.com/zoosymposia/

ISSN 1178-9905 (print edition)
ZOOSYMPOSIA
ISSN 1178-9913 (online edition)

The *Trichoptera Literature Database*: a collaborative bibliographic resource for world caddisfly research

PATINA K. MENDEZ[1,2]*, RALPH W. HOLZENTHAL[2] & JOSHUA W. H. STEINER[3]
[1]*Department of Environmental Sciences, Policy and Management, University of California, Berkeley, Berkeley, CA 94720 U.S.A. E-mail: patina.mendez@berkeley.edu*
[2]*Department of Entomology, University of Minnesota, 1980 Folwell Ave, 219 Hodson Hall, St. Paul, MN 55108 USA, E-mail: holze001@umn.edu*
[3]*Josh Steiner Consulting, 6025 Claremont Ave., Oakland, CA 94618 USA., E-mail: josh@eds.org*
(*) Corresponding author

Abstract

In addition to a list of valid names and synonyms, as provided by the *Trichoptera World Checklist*, access to the primary literature itself is essential for research in Trichoptera taxonomy and systematics. To improve access to bibliographic information, we established the *Trichoptera Literature Database,* http://www.trichopteralit.umn.edu, a bibliographic database of over 8,500 citations of literature on Trichoptera. In addition to compiling bibliographical information, we provided access to over 450 high quality Portable Document Format files (PDFs) of historically important, rare, or out-of-print older works as well as more current literature. To provide universal web access to this bibliographical resource, we constructed a dynamic, custom-designed, web application (PHP, Symfony framework) created to import Extensible Markup Language (XML) from the EndNote data file. The database allows the user to search by author and year of publication, displays citations in a standard bibliographic format, and provides download links to available PDF literature. Existing bibliographies of Trichoptera literature and online access to *Zoological Record* databases were used to accumulate citations. Protocols for scanning literature, issues regarding copyright, and procedures for uploading citations and PDFs to the database are established. We hope to create a collaborative framework of contributors by seeking regional, subject, or language organizers from the community of Trichoptera workers to assist in completing and maintaining this resource with the goal of lowering barriers to efficient access to taxonomic information.

Key words: online collaboration, historical literature, online literature, rare taxonomic literature

Introduction

An expanding number of internet resources are available to the scientific community. These resources aid communication and provide greater access to information, target areas in need of research attention, provide collaborative frameworks, and communicate science in non-traditional publishing forums. For example, taxonomic resources are being consolidated online for most taxonomic groups in the form of lists of valid species names, checklists, bibliographies, and identification tools (both traditional and interactive). Large-scale initiatives such as the *Tree of Life Web* project (http://www.tolweb.org) and the *Encyclopedia of Life* (*EOL*, http://www.eol.org) seek to

Accepted by K. Majecka, J. Majecki & J. Morse: 6 Dec. 2010; published: 10 Jun. 2011

catalog the world's biodiversity through both phylogenetic and taxonomic classification perspectives, respectively. Literature resources are becoming increasingly available through publisher archiving, general journal archiving services (e.g., *JSTOR*, http://www.jstor.org), and larger-scale projects such as the *Biodiversity Heritage Library* (http://www.biodiversitylibrary.org) and *Google Books* (http://books.google.com), which focus on out-of-print literature. Collectively, these electronic resources hold the potential to provide all of the tools (with the exception of the actual specimens) to facilitate taxonomic studies (Godfray *et al*. 2007) and to provide access to the foundational information that is sometimes difficult to obtain without access to world-class libraries.

Most scientific literature is primarily accessed online (Bollen *et al*. 2009), however the scientific literature that forms the raw material of taxonomy (Godfray *et al*. 2007) still remains largely unavailable in a digital form and difficult to access. In the case of Trichoptera, much important taxonomic literature occurs in historic, rare, and out-of-print publications, such as out-of-print foreign-language journals with limited distribution and specialized original publications (e.g., Milne 1934–1936, where each issue was distributed personally as a hand-printed copy on cropped letterhead stationery from the Harvard Museum of Comparative Zoology, John Morse, pers. comm.). Because this material is infrequently accessed and is of historic importance, libraries often store it outside of the normally circulating print material and require special permissions before granting access. For researchers without access to substantial or world-class university libraries either directly or through collaborators, the process of embarking on projects that require this foundational literature can be problematic, if not impossible. These barriers are especially high for researchers and students at small institutions, public or private agencies, or in developing countries.

Within the community of Trichoptera researchers, a number of electronic resources, primarily taxonomic in nature, are available. For example, the *Trichoptera World Checklist* (http://entweb.clemson.edu/database/trichopt/, Morse 2010) maintains a record of valid species names and associated citation information for the described species of caddisflies and serves as the taxonomic backbone of Trichoptera species names for a number of larger online resources such as the EOL. *Trichoptera Africana* (http://www.senckenberg.de/trichoptera/) provides species-level diagnostic illustrations extracted from the published literature, checklists of species, and keys to families for Western Palaearctic and Afrotropical Trichoptera. Although a number of bibliographic resources have been published, such as the *Trichopterorum Catalogus* (Fischer 1960-1973) and *Bibliographia Trichopterorum* (Nimmo 1996), no complete electronic bibliography of Trichoptera exists online; access to downloadable, digitized literature is largely confined to publisher sites and only partial representation exists within the *Biodiversity Heritage Library.* The historical literature as it pertains to Trichoptera still remains difficult to access on the internet.

In this article, we introduce the *Trichoptera Literature Database* (*TLD*, available at http://www.trichopteralit.umn.edu, Holzenthal *et al*. 2009), a collaborative resource for world caddisfly research. In creating and populating the *TLD*, we had the following objectives: (1) to compile a complete bibliographic database of existing scientific literature related to caddisflies spanning a number of subdisciplines (e.g., taxonomy and systematics, life history, biology, etc.); (2) to provide a website with search capabilities to locate bibliographic content; (3) to provide downloadable "Portable Document Format" files (PDFs) of digitally scanned literature, currently available in the public domain and targeted towards historic, rare, out-of-print, or otherwise difficult-to-access works; (4) to provide links to publisher-hosted literature available on the internet; and (5) to provide protocols and opportunities for the community of Trichoptera researchers to participate by contributing content in the form of citations and scanned literature.

History of the project

The concept for the project as a complete bibliographic database of Trichoptera initially began as a personal *EndNote* (Thompson Reuters) database for the management of the reprint collection of R. W. Holzenthal in 2000. In 2001, Holzenthal provided, by way of the University of Minnesota Insect Collection website, the *Bibliography of Neotropical Trichoptera* (no longer available, superseded by the *TLD*) as a downloadable *EndNote* file format, an interactive and searchable online *RefWorks* (RefWorks COS) database interface, and a document-formatted bibliographic list (.rtf). Most recently, this bibliography has expanded to the *Trichoptera Literature Database* (Holzenthal *et al*. 2009) to include all published caddisfly literature. The concept of a comprehensive database of Trichoptera literature was inspired also by the works of Fischer (1960-1973) and Nimmo (1996). However, in addition to providing bibliographic content, the *TLD* is intended to provide access to PDF files of historic and out-of-print literature, similar in content to *Ephemeroptera Galactica* (http://www.famu.org/mayfly/), but also to include search functionality similar to the *William L. Brown Jr. Digital Library, Database of Ant Taxonomic Literature* (http://ripley.si.edu/ent/nmnhtypedb/wlb).

Content and design

Citations and PDF files represented

The goal of this database project is to provide a comprehensive bibliography of all literature treating Trichoptera, but especially the literature on taxonomy, systematics, distribution, biogeography, and evolution of extant and fossil species. To achieve this goal, the literature is added both retroactively and proactively. Several bibliographic resources are being checked for retroactive capture of literature, primarily *Zoological Record*, but also numerous, historically published bibliographies (e.g., Betten 1934; Fisher 1960–1973; Nimmo 1996; Flint *et al*. 1999) and the literature lists compiled in *Braueria* (formerly *Trichoptera Newsletter*) and *Current and Selected Bibliographies* of the North American Benthological Society, among others. The *TLD* does not integrate specific information from within the text of the literature with the exception of abstract and keyword information from *Zoological Record*, and the *TLD* is not intended to replace works that abstract and cross-reference the literature (e.g., Fisher 1960-1973; Nimmo 1996).

Primary focus is on literature pertaining to taxonomy and systematics and related disciplines (e.g., distribution, biogeography, morphology, etc.), but we also include literature related to basic and applied ecology, behavior, biology, and physiology of caddisflies. Literature of a general ecological nature, where caddisflies are not the primary subjects of the reference, is generally not included nor is that on toxicology. Deciding on which literature or subjects to include in the database is unavoidably subjective and is ultimately at the discretion of the senior editor of the database (R. W. Holzenthal). Users who wish to include references or subjects not included in the current database are requested to send the omitted reference to the senior editor for consideration and inclusion in future versions of the database.

Digital scanning of literature is prioritized towards works that are already available in the public domain, especially taxonomic studies that are in out-of-print or difficult to access journals.

Trichoptera Literature Database strategy and structure

The *TLD* is a web-accessible application designed to provide the following services: (1) import a bibliographic XML (Extensible Markup Language) file created by the commercially available bibliographic management software, *EndNoteX2* (Thompson Reuters), into a web-accessible,

MySQL database; (2) associate and manage PDF metadata for literature available online; and (3) provide access for users in the form of webpage search forms with links to downloadable PDF content.

The primary citation database is maintained by R.W. Holzenthal as an *EndNote* file to manage the citation information (Fig.1). We chose to maintain the database in the native *EndNote* format to preserve features in *EndNote* that integrate with word processing software such as the ability to automate bibliographies and use predefined format for citations. The *EndNote* database is exported in an XML file format and is then imported by our custom database application. In addition, a copy of the literature database is provided for community users in 3 downloadable formats on the website: (1) in the native *EndNote* file format (.enl), (2) the more general RIS (.txt) format for import into other bibliographic managers such as *Zotero* (http://www.zotero.org) or *RefWorks* (http://www.refworks.com), and (3) a document format (.rtf) as a complete bibliography in the reference format of the journal *Zootaxa* (http://www.mapress.com/zootaxa/).

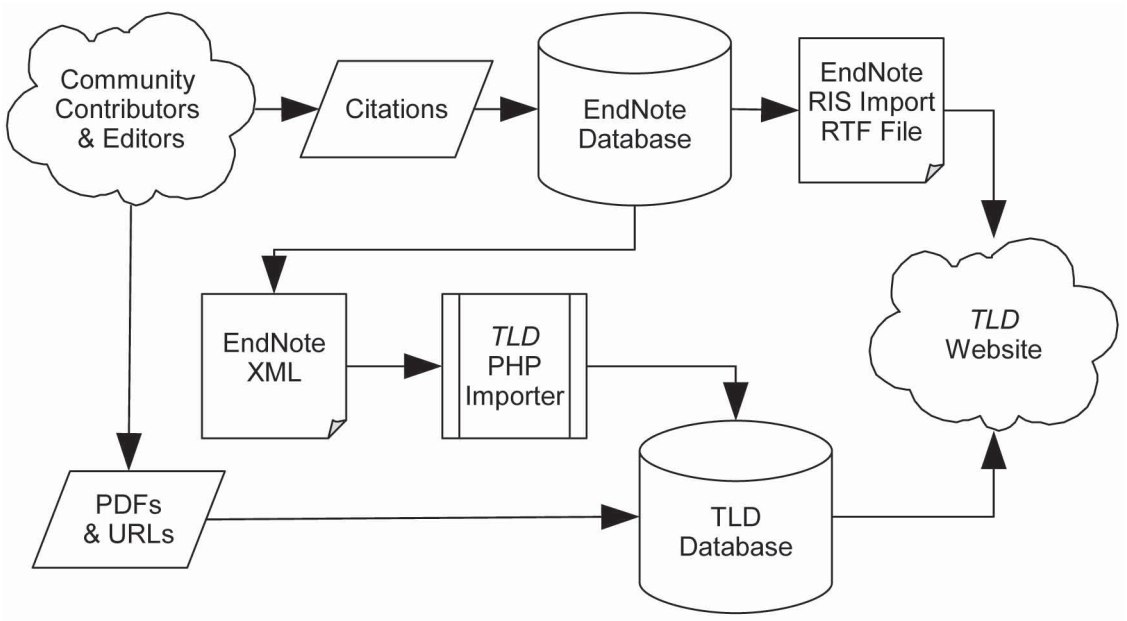

FIGURE 1. Database structure and pathways of data management. Citation information is managed in the EndNote database, exported as XML, and imported via a PHP importer into the TLD. Downloadable literature (PDFs and URLs) and linked after import.

The online application is a custom application programmed in PHP v5.2.x using the Symfony v1.2.x framework (http://symfony-project.org) and the data are stored in a MySQL database. Database records are imported from the *EndNote* file and are associated with a server-side database table that maintains information related to the PDF files. The application supports alphabetical browsing by senior author name, decade, and search functionality through web forms that allow articles to be searched by: (1) author, (2) title keyword strings, (3) publication year, (4) publication type, and (5) if a PDF file is available of the work. Search results are returned sorted by senior author name, year, and article title with 20 results per page and are formatted by the application in the *Zootaxa* reference format. Links to available PDF content are provided alongside the citation.

PDF files are managed through an administrative interface that stores the locations of (1) files hosted at the University of Minnesota and (2) files located at publisher sites. We chose to host only material that is available in the public domain, published prior to 1923 (see http://www.copyright.umn.edu/laws.html for more information on copyright laws). For literature that is

already scanned and available online, such as material on a publisher's website, or electronically archived, we provide links to the static URL of the website or to the abstract page containing information for the citation. This method of providing links to the literature stored on the publisher site allows us to avoid possible issues related to copyright restrictions in distributing the literature. Publisher-hosted material is a mix of "open access" material in which articles do not require payment to a subscription or a service and material that is restricted, requiring a subscription to the journal or service, such as a professional society membership or university library subscription. These restrictions are noted for links to publisher-hosted content.

Usage statistics are tracked through 2 mechanisms. Using *Google Analytics* (http://analytics.google.com) we compile reports of the number of page requests, the geographic location of the origins of page referrals (e.g., city, country), and date that the site was accessed. These metrics provide an assessment of the degree of worldwide access to the *TLD* as well as temporal patterns of usage of the *TLD*. In addition, from within the *TLD* application, we retain information related to search strings, links clicked to browse articles, and the articles downloaded. Tracking of user searches and metrics related to downloaded articles will allow us to focus our future scanning priorities.

Community contributions and resources

Community collaboration of this shared internet resource is essential to expand the current content to a larger digital library of downloadable literature. We ask for the following contributions with increasing levels of commitment: (1) validation of current citations for missing citations and errors and submission of PDF files (or links to journal homepages) of the collaborator's own publication record; (2) a commitment to scan and submit a body of literature for a specific author, region, or taxonomic group; and (3) serve as a "regional," "taxonomic," or "language" coordinator. Currently, contributions may be submitted via email to the editors; however, we are in the process of developing web interfaces for submission by community collaborators.

Scanning guidelines for literature are described on the website where we include a printable scanning instruction guide, and examples of acceptable and unacceptable scans. Briefly, we request that documents be saved in the Adobe Acrobat 9.0 format. Documents should be scanned in Black and White at 600 dpi (default threshold 110) and assembled into 1 complete document (or 2 or more if the work is especially large). Pages should be cropped to eliminate black margin binding areas from scanning bound materials and other stray marks (but page numbers should be visible). Pages with gray-tone photograph images should be scanned in 8-bit grayscale, 300 dpi, and pages with color illustrations or photographs should be scanned in 8-bit color, 300 dpi (file will be very large).

Project and usage statistics

The *Trichoptera Literature Database* website was publicly launched on May 26, 2009 with a little over 7000 citations. Currently the database includes over 8,500 citations and a total of approximately 10,000 citations are estimated to exist in the published literature. The database currently hosts 461 PDF files (344 scanned PDFs hosted by the *TLD* and 117 links to PDFs on publisher websites). Approximately 300 additional works are scanned, but not yet uploaded. Of the literature represented within the bibliography, we estimate that approximately 900 works were published before 1923 and can be scanned and made available through the *TLD* website. In addition, a number of works are free of copyright restrictions and will be included in our scanning priorities as well (e.g., U.S. federal government publications, uncopyrighted state government resources, reports, etc.). Most literature published within the last 10 years is available on publisher websites and can be linked from within the *TLD* (approximately 1500 references).

Since the public launch of the site, http://www.trichopteralit.umn.edu has been visited over 2100 times by users from 71 countries or territories and from 490 cities (accessed 17 May 2010), indicating to us that the website clearly has a worldwide utility and impact for caddisfly researchers. Most of the visits originated from within the United States, Japan, and Canada; however, a substantial number of visits have also occurred from across Europe, South America, Australia and New Zealand, and Asia. Already, over 30 community members have contributed additional citations, citation corrections, and PDF files.

Discussion

The *TLD* is a bibliographic resource that compliments existing internet resources available to Trichoptera researchers, such as the *Trichoptera World Checklist* and *Trichoptera Africana*. These resources currently exist as stand-alone websites, however there is potential for these databases to become interoperable. For example, within the *TLD*, we created unique identifiers for references that can be added to other existing databases to link back to source citations and literature available on the *TLD*. Communication among databases may also be facilitated by including unique identifiers for species, such as NamebankIDs available from the *Universal Biological Index and Organizer* website (*uBio*, http://www.ubio.org) or by creating life science identifiers (LSIDs, Bafna *et al*. 2008), that act as database keys and remain stable regardless of changes to the taxonomy or classification. Future databases focused on Trichoptera should consider standardizing database structures or adopting schema using those proposed by larger scale initiatives; the *Biodiversity Information Standards* (*TDWG*, http://www.tdwg.org/), the *Creating a Taxonomic E-science* project (CATE, Clark *et al*. 2009), and the *EOL* take this philosophy into consideration.

To grow the *Trichoptera Literature Database*, we welcome the participation and contributions of the community of Trichoptera researchers. In creating the database framework, we have assembled what we consider to be a comprehensive bibliography of Trichoptera. Although we have only just begun to upload PDF content, we have already received many contributions by community members. For those who contribute, we formally recognize their contributions on the website; however, we also propose to give scholarly credit by adding additional features to identify contributors directly and to attribute individual contributions from community members. In addition, we are looking forward to developing administrative tools for contributors who take ownership of taxonomic, regional, or language areas of the database.

Finally, the *TLD* underscores the importance of publishing research in "open access" journals. Although 900 references are estimated to be available in the public domain, this number represents only 10% of the estimated Trichoptera literature. More publications could be made available through the *TLD* if open access permissions are granted. A number of journals have digitized their publications and have made them available through journal archiving services such as JSTOR, but many of these archiving services require either personal or institution subscriptions for access. Many journals allow the author to pay for open access to publications (and in some cases, authors may do so retroactively), and we recommend open access publishing as the best alternative for making material available to all scientists, regardless of institutional affiliation. In some respects, although the cost may be higher for the author, if all researchers are willing to make their scholarly publications open access, the benefits of lowering barriers to access of information far outweigh the costs of the initial publication.

In summary, the *Trichoptera Literature Database* is a collaborative online bibliographic resource that provides a complete bibliography of published Trichoptera literature, hosts downloadable PDF

files, and links to publisher-hosted PDF files. We look forward to continue expanding this resource and to continue receiving support from the Trichoptera community in the form of contributed content.

Acknowledgements

We thank the many people who have contributed time and resources to this project, including: bibliographic contributors - Brady Richards, Andy Nimmo, Mike Hubbard; technical contributors: Roger Blahnik, Jolanda Huisman, Desiree Robertson-Thompson, Lourdes Chamorro, Aysha Prather, Anne Wasmund, Alex Eagen, Joel Gardner, Robin Thomson, George Guenthner; community contributors and volunteers - Brian Armitage, Tatiana Armitage, Mikhail Beketov, Clara Bicchierai, Núria Bonada, Cecilia Brand, Fernanda Cianficconi, Stanislaw Czachorowski, Ferdy de Moor, Edward DeWalt, Marianne Espeland, Gilbert Fuentes, Gísli Gíslason, Sophie Gombeer, Bert Higler, Tomiko Ito, Kjell Arne Johanson, Naotoshi Kuhara, Omar Lodovici, Arnold Móra, John Morse, Stephen Moulton II, Takao Nozaki, Carita Nybom, Ayuko Ohkawa, Andy Rasmussen, Dave Ruiter, Denes Schmera, Brian Smith, Takaaki Torii, Mariusz Tszydel, Lianfang Yang, Carmen Zamora-Muñoz; systems assistance - Michael Glaser, Shuping Zhang; and library assistance - Linda Eels, Margaret Borg, Ann Rojas.

References

Bafna, S., Humphries, J. & Miranker, D.P. (2008) Schema driven assignment and implementation of life science identifiers (LSIDs). *Journal of Biomedical Informatics,* 41, 730−738.

Betten, C. (1934) The caddisflies or Trichoptera of New York State. *New York State Museum Bulletin,* 292, 1−576.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M.A. & Balakireva, L. (2009) Clickstream data yields high-resolution maps of science. *PloS ONE,* 4, 1−11.

Clark, B.J., Godfray, H.C.J., Kitching, I.J., Mayo, S.J. & Scoble, M.J. (2009) Taxonomy as an eScience. *Philosophical Transactions of the Royal Society A*, 397, 953−966.

Fischer, F.C.J. (1960–1973) *Trichopterorum Catalogus, volumes 1-15 and index volume.* Nederlandsche Entomologische Vereeniging, Amsterdam.

Flint, O.S., Jr., Holzenthal, R.W. & Harris, S.C. (1999) *Catalog of the Neotropical Caddisflies (Trichoptera)*. Columbus, Ohio, Special Publication, Ohio Biological Survey, 239 pp.

Godfray, H.C.J., Clark, B.R., Kitching, I.J., Mayo, S.J. & Scoble, M.J. (2007) The web and the structure of taxonomy. *Systematic Biology,* 56, 943−955.

Holzenthal, R.W., Mendez, P.K. & Steiner, J.W.H. (2009) *Trichoptera Literature Database: a collaborative bibliographic resource for world caddisfly research.* Available from: http://www.trichopteralit.umn.edu (17 May 2010).

Milne, L.J. (1934–1936) *Studies in North American Trichoptera, volumes 1-3.* Privately printed, Cambridge, Massachusetts.

Morse, J.C. (Ed.) (2010) Trichoptera World Checklist. Available from http://entweb.clemson.edu/database/trichopt/ (accessed 27 May 2010).

Nimmo, A.P. (1996) *Bibliographia Trichopterorum: a world bibliography of Trichoptera (Insecta) with indexes, volume 1. 1961−1970.* Pensoft Publishers, Bulgaria, 583 pp.